Leave-Multiple-Out Informal Benchmarking

Understanding the Behavior of Informal Benchmarking for Multivariate Confounding

Author: Nayden Borodjiev
Email: N.T.Borodjiev@student.tudelft.nl
Supervisors: Jesse Krijthe, Matej Havelka

TU Delft

1. Background

Causal inference aims to estimate **cause-and-effect relationships** from data, not just correlations. In observational studies, this relies on **unconfoundedness**: all variables affecting both treatment and outcome must be observed. Hidden confounders violate this assumption and can bias estimated treatment effects.

Sensitivity analysis asks how strong such hidden confounding would need to be to change the conclusion. The Marginal Sensitivity Model captures this with **Γ**, which bounds how much an unobserved confounder can change treatment odds between individuals with the same observed covariates:

$$\Gamma^{-1} \leq \mathrm{OR}(e(x,u), e(x)) \leq \Gamma, \qquad \mathrm{OR}(a,b) = \frac{a/(1-a)}{b/(1-b)}.$$

2. LMO Informal Benchmarking

Informal benchmarking makes Γ easier to interpret by asking:
What if some observed covariates had actually been hidden?
Leave-multiple-out benchmarking drops a set S of covariates, with $|S| = m$, refits the propensity model without them, and compares the resulting treatment odds with the full model:

$$\hat{\Gamma}^{(m)} = \max_{S:|S|=m} \max_i \max \left(\frac{O_{\text{full}}(X_i)}{O_{\text{red}}^{(-S)}(X_i)}, \frac{O_{\text{red}}^{(-S)}(X_i)}{O_{\text{full}}(X_i)} \right).$$

This gives a **sample-realized benchmark**: the largest treatment-odds shift found in the observed data after hiding m covariates.

3. Research Questions

- How does the empirical leave-multiple-out benchmark behave as the number of dropped features increases?
- How does sample geometry affect the reliability of LMO benchmarking?

4. Simulation Setup and Benchmarks

- $p = 10$ bounded covariates: $X_{ij} \in [-1, 1]$.
- Treatment assignment: $e(X_i) = 1/(1 + \exp(-X_i^\top w))$.

For each omitted subset size m, the **theoretical ceiling** is

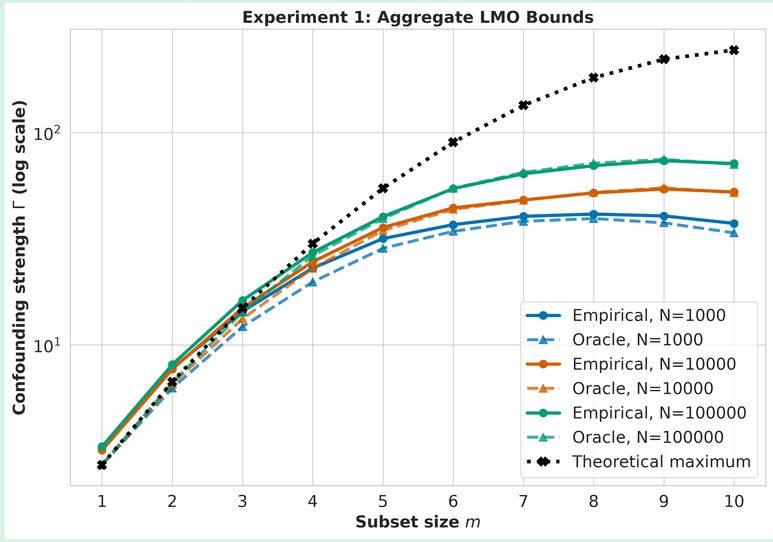
$$\Gamma_{\text{theory}}^{(m)} = \exp\left(\sum_{k=1}^{m} |w|_{(k)} \right).$$

It is the largest possible odds shift over the full bounded covariate space, so it increases monotonically with m.
The **Oracle benchmark** uses the known treatment rule, but only the observed sample:

$$\Gamma_{\text{Oracle}}^{(S)} = \max_i \exp\left(\left| \sum_{j \in S} X_{ij} w_j \right| \right).$$

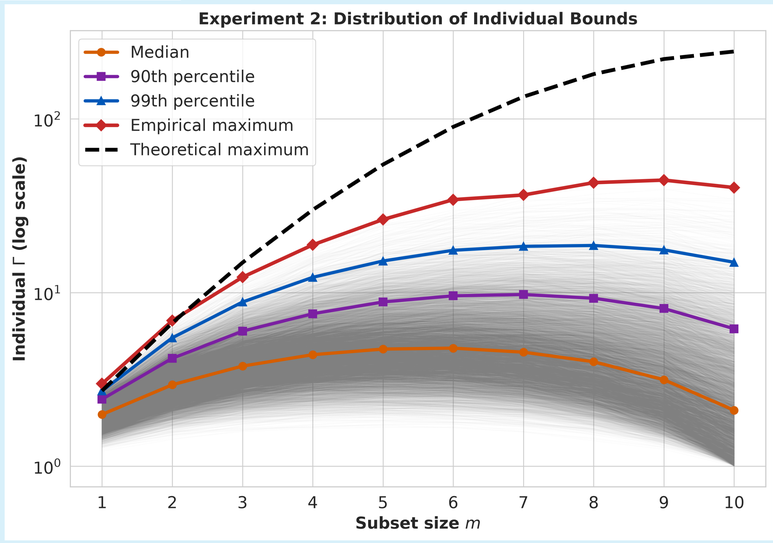
Comparing empirical, Oracle, and theoretical bounds separates **estimation error** from **sample-geometry limits**.

5. The Plateau



The theoretical Γ grows monotonically with omitted subset size because each omitted covariate can add more hidden-confounding strength. In the data, however, IB only sees the strongest odds shift among observed units. As m grows, the covariate alignments needed to reach the theoretical maximum become rare, so both empirical and oracle bounds **plateau** below theory.

6. Individual Benchmarks

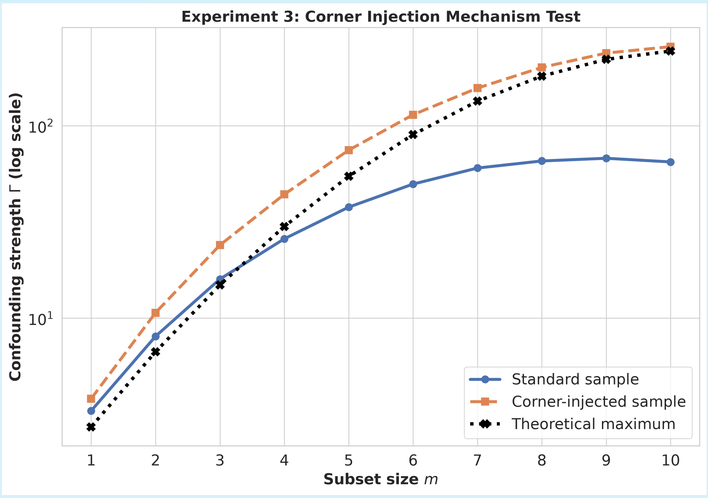


For an omitted subset S, the individual dropped log-odds contribution is

$$\Delta_i(S) = \sum_{j \in S} X_{ij} w_j \qquad \Gamma_i(S) = \exp(|\Delta_i(S)|).$$

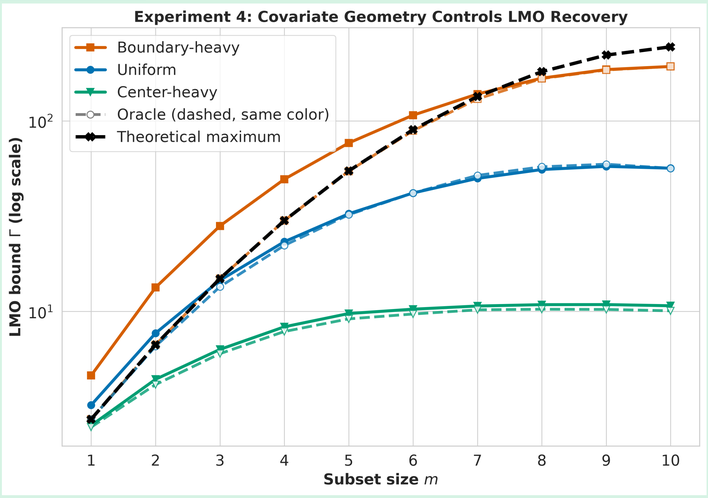
Large individual bounds occur only when omitted covariates have large magnitudes and **aligned** signs. Mixed signs cancel inside $\Delta_i(S)$, so most individuals remain far below the theoretical ceiling. The benchmark is therefore driven by **rare boundary observations**, not by the typical sample member.

7. Corner Injection



Replacing two observations with the positive and negative corners moves the benchmark closer to the theoretical curve. This supports the explanation that large LMO bounds require observed individuals with extreme, directionally aligned covariates.

8. Covariate Distribution



Changing only the covariate distribution changes the realized benchmark. Boundary-heavy samples produce larger bounds because they contain more extreme covariate patterns, while center-heavy samples under-recover the theoretical ceiling.

9. Conclusion

Main message: LMO informal benchmarking is a **sample-realized benchmark**, not a worst-case bound over the full covariate space.

- For small omitted subsets, LMO IB can provide a useful calibration of hidden confounding.
- For larger subsets, LMO IB may miss rare extreme alignments in the sample, understating sensitivity and making robustness appear stronger than it is.